18
Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge

Dominic Oldman, Martin Doerr, and Stefan Gradmann

> Meaning cannot be counted, even as it can be counted upon, so meaning has become marginalized in an informational culture, even though this implies that a judgment – that is, an assignment of meaning – has been laid upon it. Meaning lives in the same modern jail which houses the soul, the self, the ego, that entire range of things which assert their existence continually but unreasonably. (Pesce, 1999)

This chapter discusses the Semantic Web and its most commonly associated cogwheel, Linked Data. Linked Data is a method of publishing and enabling the connection of data, while the Semantic Web is more broadly about the meaning of this information and therefore the significance and context of the connections. They are often thought of as being synonymous, but the use of Linked Data in practice reveals clear differences in the extent to which the Semantic Web is realized both in terms of expressing sufficient meaning (not just to support scholarly activity but also interesting engagement) and implementing specific strategies (Berners-Lee *et al*., 2001).

These differences, particularly reflected in the approaches and outputs of different communities and disciplines operating within the humanities, bring to the fore the current issues of using Semantic technologies when working with humanities corpora and their digital representations. They also reflect more deep-seated tensions in the digital humanities that, particularly in the Open Data world, impede the formation of coherent and progressive strategies, and arguably damage its interdisciplinary objectives. We make the case for consistent forms of knowledge representation across all humanist scholarly activities correctly reflecting humanist discourse and epistemology. We also discuss the significant role that structured data, much of which has been

A New Companion to Digital Humanities, First Edition. Edited by Susan Schreibman, Ray Siemens, and John Unsworth.

contributed by humanists employed within memory institutions and recorded in institutional information systems for the last 30 years, can potentially have in the open environment of the Semantic Web. These sources have been largely overlooked as a significant source for analytical humanities research[1] (Prescott, 2012), but could provide valuable and unique meaning, context and perspective, at both micro and macro levels of research.

If the digital humanities are the "intersection between humanities scholarship and computational technologies" (Pierazzo, 2011), then Linked Data and the Semantic Web could be seen as representing polarized viewpoints from these two disciplinary cultures. As they race forward towards this imagined intersection they may either combine in a fascinating development of digital infrastructure, computer reasoning, interpretation, and digital collaboration, or instead participate in a dismal collision, leaving only a mechanical meaningless shell in its wake. Linked Data "is not enough for scientists" and therefore is not enough for humanists, and "publishing data out of context would fail to respect research methodology nor would it respect the flow of rights and reputation of the researcher" (Bechhofer *et al.*, 2013). This should apply throughout the research life cycle.

The World Wide Web sets humanists up with an almost cruel challenge. It hosts huge amounts of information about the world and its history which is increasingly difficult for the traditionalist to ignore. On the surface it provides an accessible and friendly environment for most non-technical users to browse and explore, and exerts an unquestioning acceptance about its place in the world. But as soon as we attempt to assert academic integrity onto it we find exactly the same pre-Web issues (Unsworth, 2002),[2] except they are magnified and more complicated. The options are to abandon scientific approaches and convince ourselves that the advantages of quantity and the initial accessibility of the Web of Data outweigh the concerns of loss of control, provenance, transparency, reproducibility, and all the other elements of good research (and believe that perhaps technology will sort it out later), or accept that to build a Web that truly supports the development of humanities knowledge means not accepting technology as it is served up to us, but asserting ourselves and our disciplines onto it and its development.

Linked Open Data and the Semantic Web?

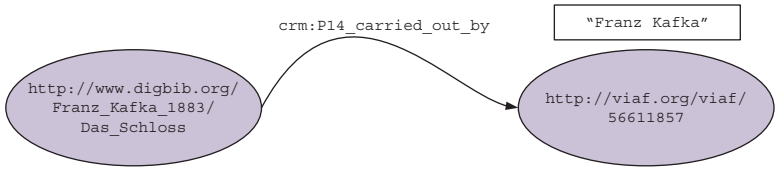
> The Web is more a social creation than a technical one. I designed it for a social effect – to help people work together – and not as a technical toy. (Berners-Lee and Fischetti, 2008)

The Semantic Web, it is argued, is the Web of meaningful data that can be processed by computers and employs "Linked Data" as the mechanism for publishing structured data to the World Wide Web where that data can be linked and integrated. It uses the same HTTP protocol (Hypertext Transfer Protocol) and a similar way of identifying data (Uniform Resource Identifiers [URI] or "web resources"), as that employed by web pages (W3C Technical Architecture Group, 2001).[3] However, in contrast to an HTML (HyperText Markup Language) Web page, the Web of Data uses a simple meta-model called RDF (Resource Description Framework) consisting of only three

elements: a subject, a predicate, and an object, commonly known as a "triple."[4] An example of such a triple statement would be:

Subject: "http://www.digbib.org/Franz_Kafka_1883/Das_Schloss"
Predicate: "http://www.cidoc-crm.org/rdfs/cidoc-crm#P14_carried_out_by"
Object: "http://viaf.org/viaf/56611857"

(the last element could also be the literal value "Franz Kafka"), or again rendered graphically:



Such triples can be combined into large, sophisticated graph structures which can be organized using a "grammar" written in the RDF Schema (RDFS)[5] language, which includes constructors for declaring sub and super classes and properties. It also incorporates the concept of inheritance, enabling simple, deterministic logical operations on such aggregations of RDF triples ("reasoning").

The Semantic Web has strong alignment with knowledge representation (a way of representing the real world designed for interpretation by computers), but information can be published as Linked Data that provides very little scope for meaningful interpretation. The clarion call from Tim Berners-Lee for open data publication has been promoted with a priority on "raw data now," with few additional public qualifications (Berners-Lee, 2009). Since the use of RDF does not mandate that data has an unfettered open license, Linked Open Data has a particular significance. If computers, rather than humans, are following and exploring links, then licensing restrictions create barriers and complexity limiting the ability to exploit the full benefits of Linked Data and Semantics – one of the main challenges cited by John Unsworth (2006) in establishing digital infrastructures. Therefore the Web of Data goes hand in hand with campaigns to change the nature of data publication to an open model that supports the advancement of more progressive knowledge objectives and outweighs the current restrictive business models entrenched in the existing Web of Pages (Renn, 2006).

The use of RDF solves substantial data integration issues by addressing the problem of schema mismatch (information modeled in different structures) and providing a platform for potentially resolving differences and equivalences in semantics. These problems of mismatch are present in other types of data model, most notably those used in relational databases and in Extensible Markup Language (XML), a format well known to many humanists as the model used for the Text Encoding Initiative (TEI).[6] The most common system of data management, relational databases, use related (or joined) tables of fields (usually highly normalized) together with a set of constraints. The associated management systems (relational database management system, RDMS) employ standards

for data query and retrieval,[7] but differences between vendors, together with different data models (different fields and structures) used for similar information mean that in practice they are unsuitable for large-scale Open World data integration. In particular, it is not possible to effectively embed the semantics of data into the underlying models.

Despite strong examples of the use of relational databases in the digital humanities, particularly in the area of prosopography,[8] lack of syntactic and semantic interoperability has inevitably limited the ability of structured data projects to reach beyond relatively narrow scopes, and has arguably contributed to a fragmentation of information and an accumulation of siloed (even if "linked") data repositories. The use of XML has provided some answers to the problem of data sharing (and is still dominant in this role) through a common and open syntax with a flexible and extensible structure. However, XML also does not address the issue of semantic interoperability and does not effectively encode meaning and relationships even within agreed schemas. Its main advantages of flexibility and extensibility create sustainability problems in that any small changes can easily break systems dependent on data integration, requiring potentially expensive ongoing maintenance and creating a constant and unacceptable risk of instability.

RDF also has its problems, but it differs in that the model is consistent across all implementations (the three main elements of the model – subject, predicate, and object – are fixed) and therefore syntactically it cannot break regardless of the information that is encoded within. Of particular significance to humanists is that semantics can be embedded (rather than described separately) within exactly the same structure. This provides far greater potential for integrating vast repositories of data using the standard Web protocol, and provides the foundation for additional technology layers with increasingly sophisticated levels of expressivity. It also provides the type of flexibility that researchers require to quickly incorporate new information and data structures that are necessary as their research progresses, and creates the opportunity for consistent forms of knowledge representation for all research activities.

The RDFS defines triples with special meaning that provide the basic building blocks for implementing hierarchical ontologies. Ontologies, in the computer science sense, are used to represent knowledge and employ poly-hierarchical structures of classes and properties reflecting different levels of specificity (or levels of knowledge) from which inferences can be made. Of particular importance for information integration is the distinct capability of RDF to formulate specialization/generalization relationships between properties or "data fields." The Web Ontology Language (OWL),[9] which really refers to a number of different implementations of knowledge representation logic, provides additional support for varying degrees of automated computer reasoning, alongside other systems,[10] to define computable relationships between concepts of different provenance.

The Semantic Web provides both short-term and long-term challenges for humanists in promoting a more meaning-orientated approach to data representation. In order to handle the tools of knowledge representation, humanists, Linked Data software developers, and infrastructure owners must develop an understanding about what kind of meaning humanists need that can be represented in these tools, and what it requires to express this meaning in terms of skills, distribution of labor, and infrastructure, for humanists and developers alike.[11] Because of a lack of deeper understanding and effective communication between these partners, and the tendency to regard technology as

the solution for self-evident applications that users "discover" and that will evolve by use on their own (Aberer *et al*., 2004), Linked Data is often seen as the finishing line without any real sense of its benefit and ultimate usefulness – it is just something that we are urged to do (Schraefel, 2007). While basic Linked Data publication may well be useful for some kinds of data, it is usually counterproductive for many types of humanities sources unless adapted to reflect specific methods and practices, and integrated into the epistemological processes they genuinely belong to.

The advanced methods of the RDF/OWL framework to express meaning and to relate and exchange it globally can only become effective if humanists engage with them and learn how to express their concepts, methods, and processes in detail, and in formalized ways. Knowledge engineering becomes a major concern in its own right. The shortcomings of the prevalent idea, that collections of intuitive lists of predicates (such as the so-called application profiles[12]) and terminology form a sufficient interface between technology and the humanists' discourse and epistemology, are reflected by the relative stagnation of developing "metadata vocabularies" and the poor results of applying reasoning methods to them, despite the continuing promises (Brown and Simpson, 2013). Humanists on their own will not be able to harness the expressive power latent in the tools without an interdisciplinary collaboration with technologists and managers in which all parties have a common understanding of the possibilities of Semantic technologies and the structure and complexity of the humanists' discourse.

Meaning and the Semantic Web

> The challenge of the Semantic Web, therefore, is to provide a language that expresses both data and rules for reasoning about the data and that allows rules from any existing knowledge-representation system to be exported onto the Web. (Berners-Lee *et al*., 2001)

Many computer scientists are familiar with Shannon's "mathematical theory of communication," which describes how information sources are encoded, transported, decoded and received in a form that is as complete and intact as possible (Shannon, 1948). Shannon assumed that the sender and receiver of a message are in perfect agreement on the meaning of the signals used. He did not consider larger numbers of users communicating via varying symbols. While the purpose of communication is to convey some meaning, the theory simply deals with the engineering problem to which the "semantic aspects of communication are irrelevant" (Shannon, 1948:349). Therefore what Shannon's theory never attempted to address was, how is meaning derived from information? This has been described as the "information paradox," in that "how can a system process information without regard to its meaning and simultaneously generate meaning in the experience of its users?" (Denning and Bell, 2012). The explanation provided by Denning and Bell added the concept that information consists of both signs and referents, and it is the association between the two that allows recipients of new information to derive new knowledge.

While this explanation fills the gap left by Shannon, it allows us to think more clearly about the importance of this association. If the signs and referents are

ambiguous, ill-defined, and disconnected from original sources, then the value of the association in deriving knowledge is diminished. While some types of simple information carry more generally understandable signs and referents with less ambiguity, this is not true of all information, and the potential for meaning to be lost, particularly in large-scale data publication, is great. This is especially true of information consisting mostly of naming "universals" that focus on the nature and type of things – "essence." In Bertrand Russell's words, information that concentrates solely on these universals, is "incomplete and insubstantial; they seem to demand a context before anything can be done with them" (Russell, 2011:64). Just as importantly, the meaning of one piece of information is not necessarily carried in one fragment. Its meaning is informed by other information (context) around it. Therefore, not only is the context of a single statement important to understand the association, but also the context provided by intentionally (and, with data integration, unintentionally) associated information.

If information (encoded knowledge) cannot provide adequate clarity (and is divorced from other contextualizing information), then this clearly becomes a problem for any further analysis because a digital representation must first and foremost provide a faithful, understandable, and explainable representation of a source as a basis for further valid scholarly investigation. It becomes difficult to produce any useful or meaningful information, however skilled the researcher and regardless of the scholarly tools wielded, if the data has weak correspondence with its original meaning. While the location of motorway roadworks or the times of trains from King's Cross Station may require less contextual framing, information in the humanities, particularly historical information, relies heavily on meaningful context from sources with different perspectives. The lack of context in digital environments is not only problematic for scholarly methodology but also impacts on any meaningful engagement of subsequent audiences. However, much of the historical information published in quantity in the Linked Data format provides very little context and therefore includes large amounts of ambiguity and misrepresentation. This can be explained, in part, by the lack of engagement or involvement of domain experts themselves in the digital representation of their data, and their lack of knowledge about the possibilities of Semantic technologies, ultimately resulting in the dominance of the technologist at the so-called intersection of digital humanities.

Computer science also seems to underestimate the challenges of representing the dependency of data on complex contexts in humanities, and does not readily assist humanists with adequate or appropriate solutions.[13] Equally, humanists are often not aware of the complexity of their own disciplinary developments and the means to structure it (as, for instance, demonstrated by Roux and Blasco, 2004) in the Linked Data world. Consequently they do not require and encourage computer scientists to take up the issue. The more insubstantial and meaningless the information published, the more humanities scholars rightfully reject it as a legitimate scholarly resource, and the less likely it is that institutions will seriously invest in Linked Data, because of a lack of benefits it provides.

The systematic and mechanical publication of data has limited practical benefits, but in the long run it is detrimental to the promotion of the disciplinary objectives of digital humanities. In the context of the "two cultures" debate, Matthew Arnold[14]

(in the nineteenth century) warned of an impending anarchy created by a "blind faith in machinery" (Arnold, 1869:sec.934), a position that has parallels with a current blind faith in Linked Data and its "anarchic," unsustainable, and un-strategic deployment. While the digital research community express concerns, these tend to concentrate on more high-level aspects such as the mechanics and functional aspects of cyberinfrastructures, particularly the role of scholarly functions or "primitives" (see below). Despite great expertise in knowledge representation in other areas of digital humanities scholarship,[15] it is often lacking in larger Open World environments, affecting the quality and meaning of information represented.

While Linked Data has become an increasingly popular way to publish data, OWL, the mechanism that supports knowledge representation on the Web, has yet to make significant inroads, with only the simplest of features being generally implemented (Glimm *et al*., 2012). While RDF provides the basis for syntactic harmonization, it is RDFS and OWL engineering (for example) that provide the key to semantic harmonization and computer interpretation, and it is this aspect of the Semantic Web that humanists might have been expected to have expressed a particular interest and concern in. This can only happen if the meaning of the predicates, terms, and vocabularies employed are more systematically developed as humanist theories in their own right, with methodologies empirically oriented towards the inference rules of the humanist discourse, such as discussed in Gardin (1990), rather than regarding human interpretation as a "black box" (Gangemi *et al*., 2005).[16]

Modeling and the Semantic Web

> There is this constant opposition between data and text. In order to process text we have to treat it as if it were data, as if text were composed of nice measurable things like characters that can be constituted into other things like words, phrases and syntagmatic objects of various kinds, and equally when we process data we try to pretend that we're doing it in a way that's not textual … that data is self-evidently not subject to interpretation. … and I am not convinced of that. (Bernard, 2011)

Modeling was argued, in the original *Companion* (McCarty, 2004), to be a fundamental activity of humanities computing and a method shared with other established disciplines. In association with knowledge representation, it has been developed in a number of different areas of humanities research. Modeling, distinguished from a model, is the ability to simulate the effects of introducing different variables and inputs. The use of acknowledged scholarly methods demonstrates academic integrity, which is important for a new field trying to establish itself. But equally important is the need to show how activities like modeling, but also other scholarly activities, continue to be applied, generating a history of development, expansion, and growing sophistication. McCarty pointed out the advantages of using computers for modeling humanities corpora in contrast to more manual approaches. Computers provide "tractability" and "absolute consistency" in an environment in which models can be manipulated with astonishing speed, but which also satisfies the computer's and modeling's necessity for precision. This makes the creation,

management, and control of larger digital datasets, representing a wider range of knowledge, problematic (McCarty, 2004:259) – and this creates a challenge for Linked Open Data environments.

McCarty identified the importance of "a structured correspondence between the model and the artifact, so that by playing with one we can infer facts about the other" (McCarty, 2004:259). In the analysis of literature this might involve the manipulation of words and word patterns and comparing the effect of these changes between an original representation of the text and subsequent manipulated versions. To produce these different outcomes (inferences) these vocabulary manipulations should operate consistently across all versions of the model within the same overall context and within the same framework of representation. In retrospect, McCarty's ultimate dissatisfaction, primarily through the modeling of Ovid's *Metamorphoses* (McCarty, 2014; see also McCarty, 1996), included the perceived inability to model context (at a micro level) objectively: "The resultant model produced interesting results but reached an impasse when I realised that its structure was not so much incomplete as arbitrary" (McCarty, 2007) (we come back to this). The development of distant reading provides a means of identifying context more systematically but from a macro or "bird's eye" position. This is where the production of Linked Data from structured information systems can provide valuable and broader historical context at all levels.

For humanities structured data (much of which comes from the information systems of cultural or memory institutions) the issue of context is different. In most organizational systems it is generally implicit, and therefore we overlook it and mistake the data for just a list of nouns.[17] However, using the knowledge of domain experts, context can be identified and represented precisely and purposefully. Making explicit this context allows analysis at both a micro and a macro level (and many levels in between), creating a highly effective knowledge system, particularly when integrated with other data. This is an extremely important aspect of the structured data that is constantly being produced by humanists like curators, librarians, and archivists. For this type of information the relationship of correspondence is different. The use of inference and analogy is not with the artifact as source material, like the text in McCarty's example, but rather it has a more direct association with *the scholar* who produces information (which is only partially recorded in an information system) that may be categorized as expressions of knowledge that are either "known facts" (often originating from those with proximity to the artifact), or expressions that are "possibly being." When these differences are distinguished and understood, the data starts to become very useful.

Semantic Web technologies provide the architecture for working with large amounts of data containing different types of fact from heterogeneous sources even within the anarchic conditions of the Web of Data, making forms of "big data" analysis possible, but still with difficulties. Modeling to find patterns in a single work of literature, like Ovid's *Metamorphoses* (McCarty, 1996) is one thing; modeling patterns *of history* (as opposed to modeling to find particular patterns within distinct historical data in which similarities and differences may be located using computer reasoning and inference) is likely to attract far more skepticism, since no system can hope to include all relevant data and context or compare with the fact that "the computer in our heads has, or can have, historical experience built into it" (Hobsbawm, 1998:38).

The question of modeling history (with its implied ability to predict future events) from large repositories of information brings to the fore a strong implicit assumption prevalent in digital humanities, that research systems should primarily contain and manipulate representations of the subject matter of humanities studies just as, for example, mechanics in physics might create a model of how rigid objects might move around. This narrow interpretation of scope immediately provokes doubt about such an endeavor in the humanities, where regularities in the subject matter are subtle, fuzzy, or rare, and the factors of influence (disciplines, mission, history, local perspectives, and so on) are countless. Even in natural sciences and in so-called "e-science," working with models of the observed or assumed reality of ultimate interest is a quite minor part of the services information systems provide. It is possible that only the discipline of meteorological forecasting broadly focuses on continually evaluating coherent models of "reality," and history works on vastly larger timescales (*longue durée*!).

The major role information systems (that now feed Linked Data repositories) can and should play is the support of the epistemological processes, i.e., what knowledge exists, where it comes from, where it has been used, where it can be used, and where it should be used – a conclusion also reached by McCarty, who described "analytic modeling,"[18] – "to raise the epistemological question of how we know what we somehow know" (McCarty, 2007:7).[19] The information system must not be seen as a surrogate of reality bound to some sort of view or filter (the use of the term "digital surrogate" is symptomatic of this confusion). Rather, it must be seen as a platform for the "externalization of argument" (Serres, 2011) to trace how different pieces of knowledge relate and how consistent they are with a past or with categorical theories possible within the limits of all known facts.

Information modeling, rather than attempting to deal with or model unlimited facts, instead pertains to the way we observe, how and under which conditions we would accept sources and adopt belief contained in them, what sorts of sources and knowledge we would use in arguments, and which sort of reasoning paradigm we apply. The final result of any academic study in the humanities or sciences constitutes only the tip of the iceberg of fact-seeking, fact-collecting and fact-evaluating activities, along with the respective documentation.

All this epistemological flow of information needs to be managed in structured data. Done adequately, it should become a representation of a combination of human behavior acting on information – the epistemology – tightly integrated with models of the reality – the ontology – that describes reality up to the level relevant to our ability to argue about them. For instance, the difference between a water glass and a wine glass may be sufficiently modeled by relating "glass" to "function," with context of "use" and "intended use," in order to relate scholarly knowledge to it. Such a model, in which the correspondence with the scholar can more easily associate relevant contextual information within a computer-compatible format, appears to be a more relevant and a far simpler way to integrate knowledge than knowing the two contextualized terms and their specializations in all languages.

In the semantics of the structural elements, the relationships which can be expressed explicitly within Linked Data become critical to the application, much more than the world describes. Even the smallest piece of information, placed in

context, may provide the missing link needed to unlock a chain of relationships in data sourced from diverse locations. The discovery of potentially related facts through the use of a particular pattern of context allows us to debate similarities and differences which we can reuse to further infer and assert various arguments and apply other evidence.

This type of Linked Data can operate on a micro level, allowing the isolation of particular information (with its perspective and context intact), or the grouping of information to provide a macro, more distant perspective. In other words, within certain types of contextual model (such as the CIDOC Conceptual Reference Model) the micro level is never lost or distorted, it simply becomes part of a density of data that can not only supply quantitative information but also "zoom" to individual instances that provide local context. Researchers can switch between facts and arguments at different levels of knowledge abstraction.

The same principles and mindset established in more discrete digital research activities should be applied to large repositories of Linked Data, and we should not be distracted by quantity. This requires the removal of a "'two cultures"' history that implies that memory institution database systems have less value than, for example, crafted TEI-type representations (Prescott, 2012).[20] Linked Data resources become richer the more they integrate (Crofts, 2004:ii) and can provide independent or complementary contexts. They should not be seen as being in opposition or competing.

In the Linked Data world we therefore have four major issues:

1. We need to differentiate between "known" facts and "possible" facts.
2. We require a model of nested (as opposed to flat) relationships, to provide the possibility of integrating data that properly represents the scholar's knowledge.
3. We need to provide information with a description of reality to the level that allows us to participate in meaningful discourse at any level.
4. We must always be able to trace the provenance of knowledge back to the source micro-level (with its original context and perspective intact).

This was impossible in the past, and is a new "innovative" ability digital humanities can provide. By representing the implicit relationships embedded in institutional datasets, an opportunity exists to establish a knowledge base that is both rich and broad enough to fuel more sophisticated digital humanities methods supported by numerous and varied historical perspectives. Collaboration with memory institutions on this single issue of digital data curation could dramatically improve the quality of humanities research, with wide-ranging benefits for society.

Digital Humanities and the Semantic Web

> You find things by the wayside or you buy a brochure written by a local historian, which is in a tiny museum somewhere, which you would never find in London. And in that you find some odd details which lead you somewhere else, and so it's a form of unsystematic searching, which of course for an academic is far from orthodoxy, because we're meant to do things systematically. (Max Sebald: Cuomo, 2011)

Anecdotal evidence suggests that those working in more established areas of the digital humanities can be skeptical of Linked Data as a disruptive threat to established methods. The current problem of "meaning" and Linked Data inevitably leads to unbalanced comparisons on quality, as if Linked Data technology itself was responsible for poor-quality data publication or the thoroughness of an institution's data recording processes.[21] This chapter has identified some of the reasons for poor-quality outputs, but in any event these comparisons of technology are not particularly useful. Knowledge representation, independent of implementation technology, is the more important foundational step for working with computers and information. All technology formats, whether XML, relational databases, or even RDF, have advantages and disadvantages. However, the purpose of the Semantic Web is to provide support for and integrate all knowledge representation systems from different domains and communities. It "allows rules from any existing knowledge-representation system to be exported onto the Web" (Berners-Lee *et al*., 2001). It is far more productive to talk about common issues of knowledge representation and understand how these systems can be improved and information better integrated. Linked Data and the Semantic Web do not invalidate existing methods of knowledge representation, and support the concept that historical studies rely on many different contexts, both digital and non-digital. This is important in gaining the confidence of a wider range of humanities scholars.

The CIDOC Conceptual Reference Model (CRM),[22] an ontology designed originally for the cultural heritage domain, but with far more scope, provides a useful case study. The CRM came about through a realization that cultural heritage institutions represented such a wide variety of different knowledge that attempting to model or integrate this within established meta-models (relational databases, or XML, for example) would be unsustainable and semantically limiting. The creation of a "bottom-up" knowledge representation method based on a continuously harmonized hierarchy of entities and relationships solved these problems and allowed the vast variety of knowledge to be sustainably managed and integrated (Doerr and Crofts, 1998).[23] The different levels of generalization and specialization created a less complex, more compact and sustainable model, but with far richer semantics enhanced using an "event"-based approach that empirically emerged from the analysis of data structures and expert practices.

As the TEI project has developed, using an XML model, it has also experienced a problem in managing an increasing level of variability and specialization, creating both management and data-integration issues. It also suffers from a lack of support for contextual semantics. Despite differences in objectives, there are similarities between the experience of humanists working with and representing structured data, and those involved in representing and analyzing text and literature. However, it would be extremely beneficial to the digital humanities as a whole if knowledge from these two communities could be better integrated.

The issues of representation for humanists working with digital text and debates about context are summarized in a number of recent conference papers. The lack of tools for semantic markup, and early initiatives and proposals for introducing RDF based solutions, were discussed at the 2010 Digital Humanities Conference (Sperberg-McQueen *et al*., 2010). At the 2014 TEI Conference a paper pointed out that "XML is

a poor language for semantic data modeling" and proposed an extension to the TEI project to include a TEI "'ontology"' and the use of RDF and Semantic Web reasoning (OWL) tools (Ciotti and Tomasi, 2014). At the 2012 Digital Humanities Conference, bearing in mind McCarty's frustration with attempting to provide a systematic approach to markup of context at the close (micro) level, scholars challenged a suggestion that distant reading makes close reading redundant and stated that the "reality is that quantitative methods are most effective when used alongside the close textual reading that allows us to contextualize the current glut of information" (Gooding *et al*., 2012). The paper argued that quality needs the continued use of micro or close reading analysis. This last point reflects a clear tension created by the lack of correspondence in digital text techniques between macro and micro approaches, something addressed in the structured data world using ontologies like CIDOC CRM. For modernists (and critics of postmodernism such as Jameson, 1991) there is still an uneasiness when we gloss over the details of history and dehumanize our memories of events that should be remembered and discussed in a more human context.

In terms of convergence, there have been ongoing attempts to bring TEI into the Semantic Web world. This has included a proposed alignment of the CIDOC CRM ontology and TEI with the objective of promoting integration between literary and textual projects, and larger repositories of cultural heritage structured data (Eide and Ore, 2007). While TEI's context is "dependent on and anchored to the objects (texts) being modeled," and CIDOC CRM relies "on a specified model of the world" (Ciula and Eide, 2014), the addition of event-based features in TEI P5[24] (names, dates, people, and places), "designed to cover a wide variety of real-world descriptions," makes it possible both to integrate the TEI P5 tag set with the real world of CIDOC CRM (Ore and Eide, 2009) and to use contextual markup by asserting CIDOC CRM entities and relationships into text directly.

The British Museum, a major knowledge and memory institution, digitally publishes its collection using CIDOC CRM knowledge representation as the basis for supporting research environments and developing better engagement possibilities.[25] At the Digital Classicist Summer Seminar in 2014 it presented a method of tagging text (in this case the Ancient Egyptian *Book of the Dead* spells and their currently unpublished translations by Egyptologist and software designer Dr. Malcolm Mosher) using CIDOC CRM (and the CRM extension FRBRoo,[26] used for bibliographic data) and RDFa,[27] which provides the ability to insert RDF Linked Data into HTML, SHTML, and XML). This allows the *Book of the Dead* text to become part of a much wider body of contextual structured information from cultural heritage sources (perhaps from Ancient Egyptian collections but also related information from other cultures and periods), blurring the border between structured databases and textual representation, creating a model that traverses the two (Norton and Oldman, 2014). While this may not address all the objectives of a TEI implementation, it nevertheless demonstrates a powerful tool for bringing text and structured historical data together.

Slowly but surely there is a move away from technology solutions that perform badly both in terms of syntax and semantics, and a renewed debate about context and its relationship with quality research. Crucially, these approaches have the potential to lead currently separated digital humanities communities towards a more integrated mode of operation and encourage the creation of integrated systems of reusable

information that retain the different and valuable perspectives of the expert groups that created them – regardless of specialism. It also opens up the possibility of uniting and strengthening the digital humanities discipline in terms of establishing a consistent representation of argument and belief that could be used across all types of humanities corpora, supporting contextual identification at both macro and micro levels, including "unsystematic" subjective propositions (not arbitrary ones) working alongside more objective but "distant" methods. In reality, unsystematic micro methods fit the big-data paradigm just as well as more systematic macro methods, as Max Sebald, carrying on from the quote above, describes:

> If you look at a dog following the advice of his nose, he traverses a patch of land in a completely unplottable manner. And he invariably finds what he's looking for. I think that, as I've always had dogs. I've learned from them how to do this. And so you then have a small amount of material and you accumulate things, and it grows; one thing takes you to another, and you make something out of these haphazardly assembled materials. (Cuomo, 2011)

Infrastructure and the Semantic Web

> Libraries, galleries, archives, museums are the very stuff of research, its heart and soul, not infrastructures. (Prescott, 2013)

Building a digital knowledge infrastructure (also known as a cyberinfrastructure) that works for the digital humanities is a complex undertaking. The report *Revolutionizing Science and Engineering Through Cyberinfrastructure* (Atkins *et al*., 2003) was an 84-page attempt to provide a comprehensive rationale for, and description of, a digital research environment that could work for any discipline. The recommended structure consists of an architectural layer with underlying components for computation, storage, and networking; a middle layer of enabling hardware, algorithmic tools, software, and operational support; and finally a service layer with applications, services, data, knowledge, and practices. The risk for such a blueprint is its own lack of correspondence with the dynamics and reality of any particular knowledge domain.

Such an environment cannot ensure successful research, because "research infrastructure is not research just as roads are not economic activity" (Rockwell, 2010). Just as Linked Data provides syntactic integration without necessarily conveying any meaning, the general-purpose cyberinfrastructure is conceived for, but uneducated by, any specific scholarly domain requirements (including the issues of data meaning and context), with the risk that technology can "distort" the methods of research (Rockwell, 2010) and that digital research can become technology-led, an issue that has arisen again and again (Oldman *et al*., 2014).

Since the Atkins report, different flavors of cyberinfrastructure have appeared with different specialisms. Some projects (e.g., Europeana; www.europeana.eu) have focused on content, becoming known as "data aggregators" and encouraging the community to create services that build on the resources they manage (although their noncollaborative methods of harvesting data have meant compromises in quality). Others have concerned themselves with providing a framework of good methodological processes

under which individual projects might operate and encourage synergies, taking a "bottom-up" approach; others have focused on specific tools and services. Almost none have focused on quality or context issues and their long-term relationship with data providers. However, current projects, for example DARIAH (*Digital Research Infrastructure for the Arts and Humanities*; www.dariah.eu) and DM2E (*Digitised Manuscripts to Europeana*; http://dm2e.eu), have focused in part on how scholarly activities might themselves be integrated. Although the functionality of tools can be informed by defining and analyzing scholarly primitives, what are their inputs and outputs and how are they practically and meaningfully connected?

The DARIAH project, in assessing data management used in individual projects, confirmed that semantics "were for the most part left implicit in these relational databases, and were complicated further by the variety of conventions used in representing data." The Semantic Web and Linked Data were thought to have "great potential … as they allow researchers to formalise resources and the links between them more flexibly, and to create, explore and query these linked resources." Further still, "ontologies can thus act as the semantic mediator between heterogeneous datasets, enabling researchers to explore, understand and extend these datasets more productively and so improve the contributions that the data can make to their research" (Blanke and Hedges, 2013:8). Similarly for DM2E, Semantic technologies play a crucial role in bringing together (providing the semantic glue) to ensure that components and processes work together effectively with a consensus as to the basic ontology of scholarly work, formalized using Linked Data (RDF) environments. Despite this, however, the focus is still currently on "functions," "operations," and "mechanics."

The next focus of attention must, if belatedly, be the sources of information that feed these scholarly activities and, as research creates new information, the outputs that these research functions produce. Traditionally, digital humanities projects have mostly crafted their own datasets limited by the resources available to any individual project. While the research questions they addressed have been useful and informative, projects lack the ability to call upon larger repositories, despite the significant amounts of accumulated data created by the large investments in digitization on the part of memory institutions over the last 30 years. This has again led to criticisms that research projects concentrate disproportionately on the technology rather than on the content they analyze and the scope of questions they address, raising the question of whether "ever-more sophisticated online resources freed up scholars to explore new ideas, or made them slaves to the digital machine" (Reisz, 2011).

The other criticism is that digital humanities initiatives have not engaged with the wider community (Zorich, 2008). This lack of connection is understandable, since institutions and aggregators have failed to document, represent, and integrate data in ways compatible with basic research standards (Terras and Ross, 2011:92). Regardless, there seems to be a distinct reluctance to work more closely with memory institutions on an equal intellectual basis to improve quality and practices in scholarly data publication (Poole, 2013: para.23). This in turn prompts comments such as "I dislike intensely the term research infrastructure. It suggests that libraries, archives, etc., [are] somehow subsidiary to research" (Prescott, 2013).

The infrastructure problem for the humanities cannot be resolved independently of addressing the sources of knowledge. The objective of Linked Data and Semantic

technologies is to encourage digital collaboration, and "help people work together" (Berners-Lee and Fischetti, 2008). It matters not how "state of the art" a cyberinfrastructure can be made, or how well scholarly methods are defined and incorporated, if the information that these components operate on lacks sufficient meaning and context. This is as true of Open World modeling as it was for McCarty's Closed World modeling – they involve the same scholarly activities and should use the same level of detail and quality.

In the humanities domain there are two significant challenges. The first is how to maximize the potential of existing sources of information, since many organizations that provide data have, by adopting digital information systems, been using Closed World models (again, semantics are "implicit," not explicit) that were never intended to fuel the type of cyberinfrastructure that we continually attempt to build. Converting this data into something that can be used by researchers requires more than a flat mechanical extraction, but rather the engagement of the community, particularly curators, archivists, and librarians, at source to provide meaningful contextualization of data before it is exported. The second is to support the transition of these source systems into ones that are specifically designed to meet the needs of a wider Open World audience, and this implies improved digital curation (Doerr and Low, 2010).

In response to these problems, ontologies have emerged that allow memory organizations to provide a research quality representation of their "closed" data models which are compatible with the Linked Data standard and fully utilize Semantic technologies.[28] Ultimately, source organizations must be involved in encoding the meaning of their own information, using their accumulated knowledge to deliver information relevant to research and a range of other uses. The investment of large amounts of money in one-size-fits-all harvesting mechanisms, and then converting this to Linked Data, removes much of its original value and provides no correspondence to original knowledge. This seems to go against the very spirit and nature of why Linked Data and Semantic technologies were created, in which enfranchisement is a key goal.

Scholarly Primitives and the Semantic Web

> Let's assume that I download onto my computer *La critique de la raison pure*, and that I start to study it, writing my comments between the lines; either I possess a very philological turn of mind and I can recognize my comments, or else, three years later, I could no longer say what is mine and what is Kant's. We would be like the copyists in the Middle Ages who automatically made corrections to the text that they copied because it felt natural to do so – in which case, any philological concern is likely to go down the drain. (Eco and Origgi, 2003:227)

In the discussion about infrastructure we found an increasing interest in revisiting and developing Unsworth's original list of scholarly functions and activities, commonly known as the "scholarly primitives": discovering, annotating, comparing, referring, sampling, illustrating, representing (Unsworth, 2000). This original illustrative list has since been expanded by various contributions (e.g., McCarty, 2003; Palmer *et al*., 2009; TaDiRAH, 2014). Increasingly different initiatives have attempted to use the

primitives as a vehicle for defining and promoting frameworks that create the "conditions" for improved data sharing and collaboration. These frameworks are intended to provide more focus and even to inform reference models to support the processes and workflows of research projects, tools, and also infrastructures.

However, while the core scholarly primitives are useful in classifying and defining activities that researchers recognize, they provide a relatively high-level standpoint and lack overall purpose in terms of insightful research outputs. Despite attempts at defining consensual definitions of the primitives, projects nevertheless create scholarly tools with a wide variation of methodological interpretation. For example, the scholarly primitive of "annotation" has been the focus of many projects over the years and a large number of annotation tools have been produced, recent ones with Linked Data outputs. In practice the exact nature of annotation as a function will always be viewed, interpreted, and manifested differently in different projects. Creating an annotation tool that works for every researcher and project would seem an unlikely outcome. In this respect the development of research activity taxonomies starts to feel similar to the development of the many other structured data terminologies. Just as application profiles are unable to define a common set of fields that can be agreed by the community, so the primitives are unable to define a fixed set of properties which belong to them, and risk becoming a diversion to supporting epistemological processes.

However, most of the core primitives are indirectly or directly related to making assertions and the generation of new facts to be encoded as new information[29] that are part of an *implicit* argument and belief value system.[30] Researchers represent, discover, compare, sample, and so on so that they can assert new statements about the materials under analysis. While the scholarly primitives are useful to identify common modes of activity, their discussion, in isolation from the representation of research outputs and conclusions, has limited the dialog about knowledge representation at the other end of the research workflow. Without attending to the representation of the results of scholarly activity we end up in a similar position to that discussed in relation to source data and its representation on the Semantic Web, but for the outputs of research. The symptoms are the same in that the community continues to define a wider and broader scope of activities that muddy the knowledge representation waters and emphasize the variability of subject matter. The unbalanced interest in the scholarly primitives might also support this chapter's contention that we are currently unable to implement a meaningful representation of scholarly work on the Semantic Web. While we understand that Semantic technologies may provide answers to these issues, the skills and knowledge necessary to move from activity definition to knowledge representation, and make the implicit explicit, are still in their early stages.

Above, we emphasized the need for correspondence between the sources of data and the analysis and layers of new information that are created as a result of research activities. The conclusion was that the propositions that we create as part of research, if they are to be analyzed in combination with, and maintain a correspondence to, source or canonical data, must be represented using the same ontological approach (with appropriate methods of differentiation).

The ontology CRMinf (an extension of the CIDOC CRM: the specification is available from www.ics.forth.gr) is one of the first knowledge representation systems to fully implement this approach. CRMinf extends the knowledge representation

principles of the CIDOC CRM and incorporates concepts from a number of argument and belief value systems (Doerr *et al*., 2011).[31] It provides the means to assert new facts using the same Linked Data patterns (graphs) implemented in the initial representation of data, but additionally supports the explicit representation of important contextual information regarding attribution and the scientific concepts of observation, inference, and belief adoption to new scholarly assertions. Additionally, it provides the means to bring different information sources with different representation systems into a common scholarly discourse even if source data itself cannot be practically integrated. A database record, a spreadsheet, a section of text, or indeed any other type of information object can be used as a premise to conclude new beliefs and create a connected and robust discourse of argument.

Argumentation, rather than just being an attachment or add-on to scholarly discourse, becomes fully integrated into the model. Extending the same principles of knowledge representation to a researcher's assertions means that computer reasoning can be used across all facts with transparency and full academic provenance. Since argumentation theory is interdisciplinary, it provides the necessary focus and appropriate scope to bring other research activities, or primitives, together.

Conclusions

> In some form, the semantic web is our future, and it will require formal representations of the human record. Those representations – ontologies, schemas, knowledge representations, call them what you will – should be produced by people trained in the humanities. (Unsworth, 2002)

Linked Data is the technical method of linking structured data, and provides an invaluable tool for bolting together, not pages of information, but structured information. Knowledge representation and Semantic technologies provide the means of elevating Linked Data to meaningful statements by communicating the intended meaning necessary for understanding these statements and their connections, in terms of not just description, but also context and provenance. This provides a basis for delivering information capable of informing a robust epistemological approach ultimately resulting in argument and belief, for which the results of other scholarly activities, including modeling and annotation, can become part of an integrated and more collaborative endeavor.

However, many internal information systems that store relevant humanities data use technologies that do not make meaning explicit, and this makes it difficult for technologists, without help from domain experts, to understand how it should be correctly represented. While a large amount of expertise and knowledge has been developed in other areas of digital humanities, some new skills are necessary to allow humanists to operate in and influence the complexities of Open World Semantics. Until this happens, the "intersection" of the digital humanities in this growing and important area will be unbalanced and waste valuable resources. This is an uncomfortable situation for humanists who regularly campaign for higher-quality information, and at the same time feel out of their depth when confronting the Linked Data

community. This has a profound effect on the ability of the Web to become a Web of Knowledge and a place to conduct serious humanities research.

Knowledge representation (an activity independent of technology), and the Semantic Web (an environment that insists on cross-disciplinary collaboration) provide the fundamental elements of a common cyberinfrastructure in which humanists can pursue individual and specialist research but in which the divisions between different research areas can be bridged. The correct application of appropriate ontologies to the highly variable outputs of humanities sources can still be integrated without a loss of local meaning and perspectives and used as context across a far broader range of research questions. The use of ontologies such as the CIDOC CRM creates a platform for precise micro and macro analysis, which can be used as supporting context for other sources of information in many different research areas. For example, digital literary history research can be enhanced by the additional context gained through structured data from memory institutions, and vice versa.

This more integrated view of research means treating cultural organizations, archives, libraries, museums, and other relevant information system sources as a part of the Academy, and part of an overall research infrastructure that promotes data quality in both inputs and outputs, as a primary concern. Experts in these institutions are part of the humanist community, not junior partners, interested practitioners, or neutral service providers (Prescott, 2012). Knowledge representation of information should, if possible, be consistent from its production to its aggregation and integration, and throughout its analysis and the assertion of argument. The representation of argument and belief should be a fundamental focus of research environments, formalized so that it can be harmonized with, differentiated from, and ultimately influence authoritative sources (and become authoritative). This provides a new dimension to analytical data modeling activities (like semantic reasoning), which can be applied across heterogeneous datasets and, in the same process, include enriching propositions made by researchers from different disciplines and organizations.

The academic community has a responsibility to ensure that the results of their work feed back into the information systems of memory institutions, and that generations of humanities scholars are able to build on the work of others, producing a stable rather than fragmented digital legacy (McGann, 2010; Prescott, 2012). There is an ongoing responsibility to improve the development of data to include, from the start, the information about significance and relevance that is currently absent from Closed World information systems (Russell *et al.*, 2009). All stakeholders should be concerned with developing improved systems of digital curation, not just the memory institutions themselves.

While we need to apply the same duty of care to structured data sources as we do in the case of other humanities sources, we need to be careful about diverting attention to objectives that are not currently within our reach and are peripheral to the solid disciplinary development of the digital humanities. This means not expending scarce resources on "dangerous exercises in futurology which think out the unthinkable as an alternative to thinking out the thinkable" (Hobsbawm, 1998:72). Humanists still need to acquire the skills that allow a more expert and authoritative contribution to the discussion of digital and web infrastructures which are currently, and unhelpfully, dominated by computer scientists and technologists.[32] In this respect the words of John Unsworth quoted at the head of this conclusion, written well over a decade ago, remain true.

Acknowledgments

We thank Ellen Van Keer (Library of Antiquity, Royal Museums of Art and History) for her kind assistance.

Notes

1 Rather than simply a reference.
2 Issues of integrity in digital projects are discussed under the term "Charlatanism" (cf. Tito Orlandi) (Unsworth, 2002).
3 Note that the term URI encompasses web resources that include URLs or web page addresses.
4 Although most systems employ another optional field, to identify a set of triples (named graphs), making a quad.
5 An RDF-based schema that provides the basic classes and properties for defining ontologies (http://www.w3.org/TR/rdf-schema).
6 See the website of the initiative at http://www.tei-c.org/index.xml.
7 The query language for relational systems is SQL (Structured Query language), informed by ISO/IEC 9075:2011.
8 For example, the *Prosopography of Anglo Saxon England* (PASE): http://www.pase.ac.uk.
9 http://www.w3.org/2001/sw/wiki/OWL.
10 For example, SPIN (http://spinrdf.org).
11 These processes are currently being defined in the CIDOC CRM Special Interest group initiative, Synergy, which provides a reference model for collaborative data provisioning. See www.cidoc-crm.org/docs.
12 Defining, amongst other things, a set of data or metadata elements that apply to a particular application but which have little application in the humanities, where these profiles cannot be defined without misrepresentation.
13 An example may be the still hesitant technical support of reification mechanisms or Named Graphs in the Semantic Web, which can be seen as a mandatory element to represent data-related argumentation in a coherent way (Doerr *et al*., 2011). For instance, the Open Annotation Model avoided the use of Named Graphs because of concerns about their maturity, resulting in relatively complex workarounds in contrast to those presented in Serna *et al*. (2011).
14 A poet and educationalist – he debated with Thomas Huxley on the balance of culture and science in society.
15 TEI, for example, is a form of knowledge representation.
16 As opposed to "glass box," where "we can treat the internal structure of those data *as if* it is the internal structure of an expertise." The reason why ontologies like CIDOC CRM (see below) are "bottom-up" in design.
17 For example, "The 'nouns' are the pieces of data or information the user wants" (Winesmith and Carey, 2014).
18 Rather than attempting to model history.
19 McCarty lists five trajectories with the more practical at the top. "1. A world-wide, semi-coordinated effort to create large online scholarly resources; 2. Out of this activity, the slow development of new genres in something like a digital Library; 3. Analytic modelling, to raise the epistemological question of how we know what we somehow know; 4. Synthetic modelling, to reconstruct lost artefacts from fragmentary evidence, blurring gradually into a 5. Modelling for possible worlds".
20 For an example, see http://sites.tufts.edu/liam/2014/04/23/trends.
21 See LiAM (2014): an example of comparing TEI sources with Linked data from structured sources.
22 www.cidoc-crm.org – "provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation."
23 Also see the CIDOC CRM Primer at http://www.cidoc-crm.org/docs/CRMPrimer.pdf.
24 http://www.tei-c.org/Guidelines/P5.
25 A Linked Data interface at http://collection.britishmuseum.org, and ResearchSpace at http://www.researchspace.org.
26 An object-orientated ontology version of the model, Functional Requirements for Bibliographic Records.
27 See http://www.w3.org/TR/xhtml-rdfa-primer.
28 Most notably the CIDOC CRM (Conceptual Reference Model), although this, while having the ability to be implemented using Linked Data, is technology-agnostic.

29 Tools like the DM2E Pundit annotation system (http://dm2e.eu/digital-humanities) show a movement towards a full argument and belief value system.

30 An analogy to the implicit relationships in structured data information systems.

31 Includes argumentation examples from the following papers: Toulmin (2003), Kunz & Rittel (1970), Pinto *et al*. (2004).

32 For example, see the W3C Linked Open data and Semantic Web mailing lists.

References and Further Reading

Aberer, K., Cudré-Mauroux, P., Ouksel, A.M., *et al*. 2004. Emergent semantics principles and issues. In *Database Systems for Advanced Applications*, ed. Y. Lee, J. Li, K.-Y. Whang, and D. Lee. Belin: Springer, 25–38. http://link.springer.com/chapter/10.1007/978-3-540-24571-1_2 (accessed October 12, 2014).

Antoniou, G., and Van Harmelen, F. 2004. *A Semantic Web Primer*. Cambridge, MA: MIT Press. http://www.dcc.fc.up.pt/~zp/aulas/1415/pde/geral/bibliografia/MIT.Press.A.Semantic.Web.Primer.eBook-TLFeBOOK.pdf (accessed October 31, 2014).

Arnold, M. 1869. *Culture and Anarchy*. London: Smith, Elder & Co.

Atkins, D., Droegemeier, K.K., Feldman, S.I., *et al*. 2003. Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. National Science Foundation. https://arizona.openrepository.com/arizona/handle/10150/106224 (accessed September 17, 2014).

Bechhofer, S., Buchan, I., De Roure, D., *et al*. 2013. Why linked data is not enough for scientists. *Future Generation Computer Systems* 29 (2), 599–611.

Bernard, L. 2011. *Data vs. Text: Forty Years of Confrontation. Hidden Histories Symposium (UCL)*. Hidden Histories. University College London. http://hiddenhistories.omeka.net/items/show/8 (accessed August 5, 2013).

Berners-Lee, T. 2009. The next web. http://www.ted.com/talks/tim_berners_lee_on_the_next_web (accessed September 17, 2014).

Berners-Lee, T., and Fischetti, M. 2008. Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor. San Francisco: Harper.

Berners-Lee, T., Hendler, J., and Lassila, O. 2001. The semantic web. *Scientific American* 284 (5), 28–37.

Blanke, T., and Hedges, M. 2013. Scholarly primitives: Building institutional infrastructure for humanities e-Science. *Future Generation Computer Systems* 29 (2), 654–61.

Brown, S., and Simpson, J. 2013. The curious identity of Michael Field and its implications for humanities research with the semantic web. In *Big Data, 2013 IEEE International Conference on*. IEEE, 77–85. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6691674 (accessed October 10, 2014).

Ciotti, F., and Tomasi, F. 2014. Formal ontologies, Linked Data and TEI. In *Decoding the Encoded*. Evanston, IL: Text Encoding Initiative. http://tei.northwestern.edu/files/2014/10/Ciotti-Tomasi-22p2xtf.pdf (accessed October 29, 2014).

Ciula, A., and Eide, Ø. 2014. Reflections on cultural heritage and digital humanities: modelling in practice and theory. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*. New York: ACM, 35–41. http://doi.acm.org/10.1145/2595188.2595207 (accessed October 29, 2014).

Crofts, N. 2004. Museum informatics: the challenge of integration. University of Geneva. http://archive-ouverte.unige.ch/unige:417 (accessed July 23, 2014).

Cuomo, J. 2011. A conversation with W.G. Sebald (interview). In *The Emergence of Memory: Conversations with W.G. Sebald*, ed. L.S. Schwartz. New York: Seven Stories Press, 93–118.

Denning, P.J., and Bell, T. 2012. The information paradox. *American Sciientist* 100, 470–7.

Doerr, M., and Crofts, N. 1998. *Electronic Esperanto: the role of the oo CIDOC Reference Model*. Citeseer. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.47.9674&rep=rep1&type=pdf (accessed August 26, 2013).

Doerr, M., and Low, J. T. 2010. A postcard is not a building why we need museum information curators. In *ICOM General Conference, Shanghai, China*. https://www.ics.forth.gr/_publications/

CIDOC_2010_low_martin.pdf (accessed November 1, 2014).

Doerr, M., Kritsotaki, A., and Boutsika, K. 2011. Factual argumentation: a core model for assertions making. *Journal on Computing and Cultural Heritage*, 3(3), p.1–34.

Eco, U., and Origgi, G. 2003. Auteurs et autorité: un entretien avec Umberto Eco. *Texte-e: Le texte à l'heure de l'Internet*, 215–30.

Eide, Ø., and Ore, C.-E. 2007. From TEI to a CIDOC-CRM Conforming Model: Towards a Better Integration Between Text Collections and Other Sources of Cultural Historical Documentation. In *Get Swept Up In It*. Digital Humanities 2007, University of Illinois, Urbana–Champaign. http://www.edd.uio.no/artiklar/tekstkoding/poster_156_eide.html (accessed October 28, 2014).

Gangemi, A., Catenacci, C., Ciaramita, M., and Lehmann, J. 2005. A theoretical framework for ontology evaluation and validation. *SWAP*. Citeseer. http://www.loa.istc.cnr.it/old/Papers/swap_final_v2.pdf (accessed October 28, 2014).

Gardin, J.-C. 1990. The structure of archaeological theories. In *Mathematics and Information Science in Archaeology: A Flexible Framework*, ed. A. Voorrips (ed.). Studies in Modern Archaeology 3. Bonn: Holos, 7–25.

Glimm, B. Hogan, A., Krötzsch, M., and Polleres, A. 2012. OWL: yet to arrive on the Web of Data? arXiv preprint arXiv:1202.0984. http://arxiv.org/abs/1202.0984 (accessed September 14, 2014).

Gooding, P., Warwick, C., and Terras, M. 2012. The myth of the new: mass digitization, distant reading and the future of the book. In *Digital Humanities 2012, Hamburg*. http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/the-myth-of-the-new-mass-digitization-distant-reading-and-the-future-of-the-book.1.html (accessed October 29, 2014).

Gradmann, S., and Meister, J.C. 2008. Digital document and interpretation: re-thinking "text" and scholarship in electronic settings. *Poiesis & Praxis* 5 (2), 139–53.

Heath, T., and Bizer, C. 2011. *Linked Data: Evolving the Web into a Global Data Space*. San Rafael, CA: Morgan & Claypool.

Hobsbawm, E. 1998. *On History*, new edition. London: Abacus.

Hooland, S. van, and Verborgh, R. 2014. Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish Your Metadata. London: Facet.

Jameson, F. 1991. Postmodernism, or the Cultural Logic of Late Capitalism. Durham, NC: Duke University Press.

Kunz, W., and Rittel, H.W.J. 1970. *Issues as Elements of Information Systems*. Institute of Urban and Regional Development, University of California.

LiAM. 2014. Trends and gaps in linked data for archives. LiAM: Linked Archival Metadata. http://sites.tufts.edu/liam/2014/04/23/trends (accessed October 28, 2014).

McCarty, W. 1996. Finding implicit patterns in Ovid's *Metamorphoses* with TACT. *CH Working Papers*. http://journals.sfu.ca/chwp/index.php/chwp/article/view/B.3/91 (accessed December 27, 2011).

McCarty, W. 2003. Humanities computing. *Encyclopedia of Library and Information Science* 2, 1224.

McCarty, W. 2004. Modeling: a study in words and meaning. In *A Companion to Digital Humanities*, ed. S. Schreibman, R. Siemens, and J. Unsworth. Oxford: Blackwell. http://www.digitalhumanities.org/companion (accessed December 27, 2011).

McCarty, W. 2007. Looking backward, figuring forward: modelling, its discontents and the future. In *Digital Humanities 2007*, University of Illinois Urbana–Champagne. http://www.mccarty.org.uk/essays/McCarty,%20Looking%20backward.pdf (accessed October 24, 2014).

McCarty, W. 2014. Getting there from here: remembering the future of digital humanities. Roberto Busa Award lecture 2013. *Literary and Linguistic Computing* 29 (3), 283–306.

McGann, J. 2010. Sustainability: the elephant in the room. In *The Shape of Things to Come*. A Mellon Foundation Conference at the University of Virginia. http://shapeofthings.org/papers/JMcGann.docx (accessed May 18, 2014).

Moretti, F. 2007. Graphs, Maps, Trees: Abstract Models for a Literary History. London: Verso.

Nen, E.H. 2012. Publishing and Using Cultural Heritage Linked Data on the Semantic Web. San Rafael, CA: Morgan & Claypool.

Norton, B., and Oldman, D. 2014. A new approach to digital editions of ancient manuscripts using CIDOC-CRM, FRBRoo and RDFa. In *Digital Classicist London & Institute of Classical Studies seminar 2014, UCL, London*. http://www.digitalclassicist.org/wip/wip2014-10do.html (accessed October 29, 2014).

Oldman, D. Doerr, M., de Jong, G., Norton, B., and Wikman, T. 2014. Realizing lessons of the

last 20 years: a manifesto for data provisioning and aggregation services for the digital humanities (a position paper). *D-Lib Magazine* 20 (7/8). http://www.dlib.org/dlib/july14/oldman/07oldman.html (accessed July 15, 2014).

Ore, C.-E., and Eide, Ø. 2009. TEI and cultural heritage ontologies: exchange of information? *Literary and Linguistic Computing* 24 (2), 161–72.

Palmer, C.L., Teffeau, L.C., and Pirmann, C.M. 2009. Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development. Dublin, OH: OCLC Programs and Research. http://www.oclc.org/programs/publications/reports/2009-02.pdf (accessed October 13, 2014).

Pesce, M. 1999. SCOPE1: information vs. meaning. In *Hyperreal, Vienna*. http://hyperreal.org/~mpesce/SCOPE1.html (accessed October 5, 2014).

Pierazzo, E. 2011. Digital humanities: a definition. http://epierazzo.blogspot.co.uk/2011/01/digital-humanities-definition.html (accessed July 16, 2013).

Pinto, H.S., Staab, S., and Tempich, C. 2004. DILIGENT: towards a fine-grained methodology for DIstributed, Loosely-controlled and evolvInG Engineering of oNTologies. In *ECAI 2004: Proceedings of the 16th European Conference on Artificial Intelligence*, ed. R. López de Mántaras. Amsterdam: IOS Press, 393–7.

Poole, A. 2013. Now is the future now? The urgency of digital curation in the digital humanities. *DHQ: Digital Humanities Quarterly*, 7 (2). http://www.digitalhumanities.org/dhq/vol/7/2/000163/000163.html (accessed October 24, 2014).

Prescott, A. 2012. An electric current of the imagination. *Digital Humanities:Works in Progress*. http://blogs.cch.kcl.ac.uk/wip/2012/01/26/an-electric-current-of-the-imagination (accessed March 15, 2012).

Prescott, A. 2013. Andrew Prescott (@Ajprescott) | Twitter. https://twitter.com/Ajprescott (accessed October 17, 2014).

Reisz, M. 2011. Surfdom. *Times Higher Education*. http://www.timeshighereducation.co.uk/story.asp?storycode=418343 (accessed December 28, 2011).

Renn, J. 2006. Towards a web of culture and science. *Information Services and Use* 26 (2), 73–9.

Rockwell, G. 2010. As transparent as infrastructure: on the research of cyberinfrastructure in the humanities. *openstax cnx*. http://cnx.org/contents/fd44afbb-3167-4b83-8508-4e70885b6136@2 (accessed September 21, 2014).

Roux, V. and Blasco P. 2004. *Logicisme et format SCD: d'une épistémologie pratique à de nouvelles pratiques éditoriales Hermès*. Paris: CNRS-éditions.

Russell, B. 2011. The Problems of Philosophy. Vook.

Russell, R., Winkworth, K., and Collections Council of Australia. 2009. *Significance 2.0: A Guide to Assessing the Significance of Collections*. Rundle Mall, SA: Collections Council of Australia.

Schraefel, M.C. 2007. What is an analogue for the semantic web and why is having one important? *ACM SIGWEB Newsletter*, Winter 2007. http://eprints.soton.ac.uk/264274/1/schraefelSWAnalogueHT07pre.pdf (accessed September 17, 2014).

Serres, M. 2011. Interstices: les nouvelles technologies, que nous apportent-elles? *Interstices*. https://interstices.info/jcms/c_15918/les-nouvelles-technologies-que-nous-apportent-elles (accessed October 16, 2014).

Shannon, C.E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* XXVII (3). http://www3.alcatel-lucent.com/bstj/vol27-1948/articles/bstj27-3-379.pdf (accessed September 21, 2014).

Sperberg-McQueen, C.M., Marcoux, Y., and Huitfeldt, C. 2010. Two representations of the semantics of TEI Lite. In *Cultural Expression, Old and New. Digital Humanities 2010, King's College, London*. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-663.html (accessed October 29, 2014).

TaDiRAH. 2014. TaDiRAH: Taxonomy of Digital Research Activities in the Humanities. *Dariah*. http://tadirah.dariah.eu/vocab/index.php (accessed October 13, 2014).

Terras, M., and Ross, C. 2011. Scholarly information-seeking behaviour in the British Museum online collection. In *Museums and the Web 2011, Philadephia*. http://www.museumsandtheweb.com/mw2011/papers/scholarly_information_seeking_behaviour_in_the.html (accessed October 29, 2014).

Toulmin, S.E. 2003. *The Uses of Argument*. Cambridge: Cambridge University Press.

Unsworth, J. 2000. Scholarly primitives: what methods do humanities researchers have in common, and how might our tools reflect this? Paper presented at *Humanities Computing: Formal Methods and Experimental Practice*, King's College,

London. http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html (accessed October 2014).

Unsworth, J. 2002. What is humanities computing and what is not? http://computerphilologie.uni-muenchen.de/jg02/unsworth.html (accessed December 27, 2011).

Unsworth, J. 2006. Our Cultural Commonwealth: the report of the American Council of learned societies commission on cyberinfrastructure for the humanities and social sciences. ACLS: New York. https://www.ideals.illinois.edu/handle/2142/189 (accessed September 22, 2014).

W3C Technical Architecture Group. 2001. *Architecture of the World Wide Web, Volume One*. http://www.w3.org/TR/webarch (accessed September 14, 2014).

Winesmith, K., and Carey, A. 2014. Why build an API for a museum collection? San Francisco Museum of Modern Art. http://www.sfmoma.org/about/research_projects/lab/why_build_an_api (accessed November 9, 2014).

Zorich, D. 2008. *A Survey of Digital Humanities Centers in the United States*. Washington, DC: Council on Library and Information Resources. http://www.clir.org/pubs/reports/pub143/contents.html (accessed November 9, 2014).